Tabular data Initial EDA
Initial EDA basic steps

EDA - exploratory data analysis

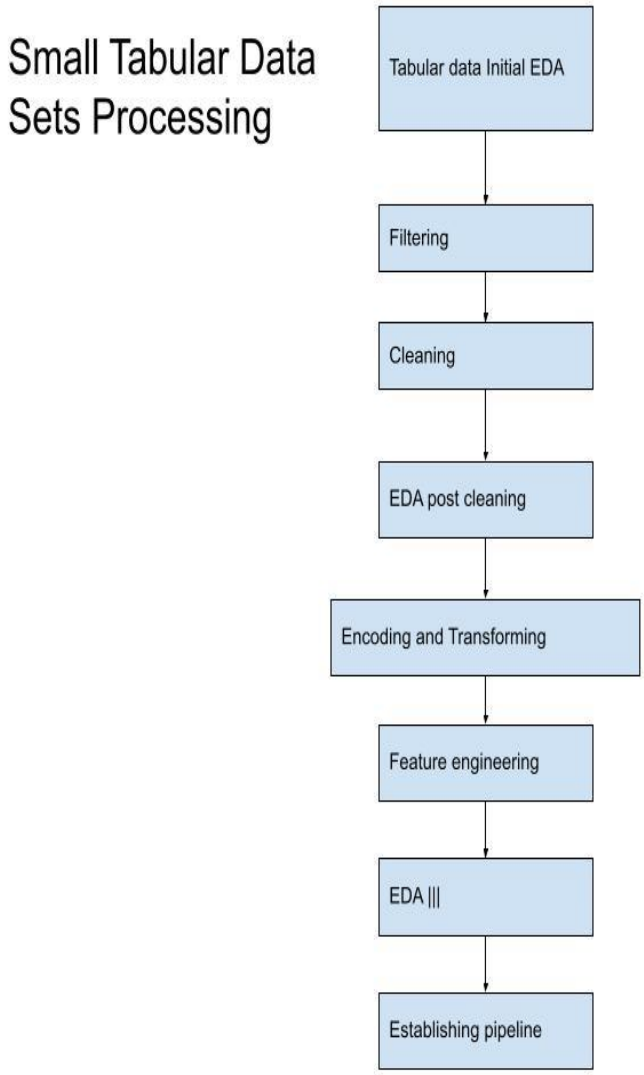
Why and What :

Why - When you receive a dataset, developing a predictive model is not akin to following a cake recipe or an instruction manual for repairing a damaged car; it is, by definition, an exploration into uncharted territory. Therefore in the first step you need to get a clearer understanding of what is your starting point before starting to navigate.
What - To comprehend the dataset, you need to perform basic operations that reveal a portion of the underlying structure, offering a direct view of the raw data, visual dependencies, and distributions, thereby understanding the story of the data firsthand.

What is it good for ? The initial EDA is a step before the preprocessing steps:

I. Filtering
II. Cleaning
III. EDA II
IV. Encoding
V. Transforming
VI. Feature engineering
VII. EDA |||

These steps will create the pipeline. Constructing this pipeline is a critical step in developing a model. Attempting to establish such a pipeline without clear insight is nearly impossible, due to the vast number of possibilities for each step. Moreover, carefully considering and documenting the reasons for choosing one step over another is crucial for creating a successful model.



The building blocks of initial EDA :

1. Uploading the data, yes people underestimating this step but is far from easy in cases there are multiple sources of data
2. Basic information of the dataframe
   - 2.1. Watching the 5 first rows, 5 last rows and random sample of the data
   - 2.2. Shape of the data
   - 2.3. Column names
   - 2.4. Finding the data types of each one of the columns
   - 2.5. Finding the number of different values for each one of the columns (for large data sets take a sample)
   - 2.6. Number of NaNs or None for each column
3. Data Quality check

3.1. Consistency of Categorical Variables: We should check for any inconsistencies in the categorical variables, such as misspelled categories or categories that should be combined.
3.2. Duplicates: It's important to check for and remove any duplicate records that could skew the analysis.
3.3. Irrelevant Records: Identify any records that don't make sense within the context of the data, such as customers with a tenure of zero but with significant total charges.
3.4. Validity: Ensure that all data entries are valid according to domain knowledge.
3.5. Accuracy: Check for data entry errors.
3.6. Completeness: We've already addressed missing values, but we should also ensure that all necessary data is present and that there are no unexpected empty fields.
3.7. Uniformity: The data should be consistently formatted. For example, ensure that all monetary values are in the same currency and format.

4. Visualizing columns data with respect to target column
    4.1. Univariate analysis :
        4.1.1. Categorical data : distribution , balanced/unbalanced
        4.1.2. Numerical data : histogram viewing the distribution of the data, looking for outliers
        4.1.3. Other types of data:
            4.1.3.1. Text data - require more involved analysis for later stages
            4.1.3.2. Time series - require more involved analysis for later stages
        4.1.4. Target column , there the following cases:
            4.1.4.1. Categorical - checking distribution with respect to the number of options
            4.1.4.2. Binary - balanced/imbalanced and if strictly imbalanced at what degree ?
            4.1.4.3. Numerical - histogram, normal/uniform/other has fat tails? skewed ?
    4.2. Multivariate analysis :
        4.2.1. Relationships between column and target column
            4.2.1.1. Categorical: relative size
            4.2.1.2. Numerical: shaded part of the histogram
        4.2.2. Faceted grids for understanding interactions between multiple variables.
        4.2.3. Use dimensionality reduction techniques like PCA to identify patterns in the data.

4.2.4. Advanced clustering methods could also reveal natural groupings within the data.

5. Segmentation analysis
    5.1.1. Perform cluster analysis to identify distinct segments within the customer base.
6. How to use the initial EDA?
    6.1. Filtering , sparse columns with limited data in small data sets (less than 10000 records) should be omitted
    6.2. Cleaning:
        6.2.1. Imputation that is completion of missing values (NaN and similar)
        6.2.2. Outliers removal
        6.2.3. Followup with inconsistent data
    6.3. Encoding of categorical data
    6.4. Transformation -> fit the transformation to the distribution